IDENTIFICATION PROBLEMS IN THE SOCIAL SCIENCES AND LIFE

Charles F. Manski
Board of Trustees Professor in Economics
Northwestern University

Università Degli Studi Di Roma 'Tor Vergata': May 24, 2006

Some Sources

Manski, C., *Identification Problems in the Social Sciences*, Harvard University Press, 1995.

Manski, C., *Partial Identification of Probability Distributions*, Springer-Verlag, 2003.

Manski, C., *Social Choice with Partial Knowledge of Treatment Response*, Princeton University Press, 2005.

1. Introduction

The Reflection Problem

Here is an identification problem. It is common to observe that persons who belong to the same group tend to behave similarly. Two hypotheses advanced to explain this phenomenon are

endogenous effects, wherein the propensity of an individual to behave in some way varies with the prevalence of the behavior in the group

correlated effects, wherein individuals in the same group tend to behave similarly because they face similar environments and have similar personal characteristics.

Similar behavior within groups could stem from endogenous effects; for example, group members could experience pressure to conform to group norms. Or group similarities might reflect correlated effects; for example, persons with similar characteristics might choose to associate with one another. Empirical observations of the behavior of individuals in groups cannot by itself distinguish these hypotheses. To draw conclusions requires combination of empirical evidence with sufficiently

strong assumptions about the nature of individual behavior and social interactions.

Why might one care whether observed patterns of behavior are generated by endogenous effects, by correlated effects, or in some other way? A good practical reason is that different processes have differing implications for public policy. For example, understanding how students interact in classrooms is critical to the evaluation of many aspects of educational policy, from ability tracking to class size standards to racial integration programs.

I have called this the *reflection problem* because it is similar to an inferential problem that occurs when one observe the almost simultaneous movements of a person and of his image in a mirror. Does the mirror image cause the person's movements, does the image reflect the person's movements, or do the person and image move together in response to a common external stimulus? Empirical observations alone cannot answer this question. Even if you were able to observe innumerable instances of persons and their mirror images, you would not be able to deduce the process at work. To reach a conclusion requires that you understand something of optics and of human behavior.

Methodological Research

Identification problems are problems of deductive logic. The conclusions that can logically be drawn depend on the assumptions and data that are brought to bear. The available data about human behavior are typically limited and the range of plausible assumptions wide. Researchers who analyze the same data under different maintained assumptions may, and often do, reach different logically valid conclusions.

Econometricians and other methodologists cannot solve this problem. What we can do is to clarify what conclusions can and cannot be drawn with various combinations of data and assumptions.

Identification and Statistical Inference

For over a century, methodological research in the social sciences has made productive use of probability and statistics. One supposes that the objective is to infer some feature of a population described by a probability distribution and that the

data are observations extracted from the population by some sampling process. One combines the data with assumptions about the population and the sampling process to draw statistical conclusions about the population feature of interest.

Working within this familiar framework, econometricians have found it useful to separate inference into statistical and identification components. Studies of *identification* determine the conclusions that could be drawn if one were able to observe a data sample of unlimited size. Statistical inference seeks to characterize how sampling variability affects the conclusions that can be drawn from samples of limited size.

My Program

Throughout my career, I have been concerned with both the identification and the statistical components of inference. My earliest published research, based on work in my dissertation on discrete choice analysis, concerned the problem of inference on people's preferences from observations of the choices that they make. Economists often say that choice behavior "reveals preferences." In fact, observation of the action that a person chooses only reveals that this action is weakly preferred to all

other feasible actions. It does not reveal how the person ranks non-chosen actions relative to one another. Thus, choice data only reveals a little bit about preferences.

Discrete choice analysis as practiced in econometrics combines data on choices with assumptions about the decision rules that persons use to make the choices that researchers observe. I have sought to understand what can be learned about decision processes given relatively weak assumptions about these processes. My earliest finding in the 1970s on identification of random utility models yielded a new method, *maximum score* estimation, for statistical inference on preferences from sample data on choices. Since then, we have learned much more about inference from discrete choice data.

Over time I have, in my research and teaching, gradually devoted less attention to the study of statistical questions and more to the analysis of identification. I have come to think that, although statistical problems contribute to the difficulty of empirical research, identification is the more fundamental problem of the social sciences.

2. Broad Themes

In several books and numerous articles, I have examined identification problems that arise in the analysis of treatment response, in studies of social interactions, in the interpretation of nonresponse in surveys, and in inference on preferences and expectations. This research has yielded many specific new findings, but I am most proud of some broad, related themes that have emerged over the past fifteen years. I have come to feel that these themes should influence the manner in which social scientists perform empirical research. These themes are:

Begin with the Data Alone

The prevalent approach to empirical research in the social sciences has been to maintain assumptions that are strong enough to fully identify quantities of interest and to yield statistically precise point estimates of these quantities. Concerns about the credibility of assumptions are addressed through the performance of specification tests and/or sensitivity analyses. Concerns about credibility are also addressed by exploring how estimates change and statistical precision falls as functional form and distributional assumptions are weakened.

A complementary approach to empirical inference begins by asking what can be learned from the data given only knowledge of the sampling process and no other information. Having determined this, one may then ask what more can be learned given successively stronger maintained assumptions. This approach yields a series of successively tighter bounds on quantities of interest. The bound is widest when no assumptions are maintained and narrows as stronger assumptions are imposed. Sufficiently strong assumptions narrow the bound to a point.

I feel that this approach to empirical research has particular value when social scientists invoking different strong assumptions find themselves in disagreement about the interpretation of empirical evidence. Establishing the conclusions that hold up under weak assumptions can build a domain of consensus, and confine disagreements to questions whose resolution really require controversial assumptions.

Inference with Missing Data: A pervasive problem of survey research is that some persons do not respond to a question posed, say a question about their income. Suppose that one wants to use the survey data to learn the population distribution of income.

Two common ways to cope with nonresponse are (a) drop the observations

with missing data and (b) impute the missing values. Both yield point estimates of the income distribution, but at the cost of nontestable assumptions. The first implicitly assumes that nonresponse is random and the second assumes whatever process is used to generate the imputations.

I recommend instead that one begin by asking what conclusions can be drawn about the distribution of income with no assumptions about the nature of the missing data. The result is a bound on the income distribution, whose width depends on the frequency of missing data. This done, one may entertain assumptions that enable sharper conclusions.

Partial Identification

I have just now referred to bounds on quantities of interest. Social scientists have commonly thought of identification as a yes/no question: a parameter is either identified or not identified. Yet identification generally is not a binary state. A researcher who does not have rich enough data and other information to infer the exact value of a parameter may nevertheless be able to partially identify it. That is, one may be able to learn that the parameter lies in some restricted set of values.

The fixation of social scientists on point identification long inhibited appreciation of the potential usefulness of bounds. I use the term "fixation" because I cannot readily understand the scientific basis for the traditional idea that a parameter is either identified or not. Bounds on parameters have been reported from time to time in the methodological literature. Nevertheless, in empirical research and in the teaching of econometrics, identification has generally been thought of as point identification.

The Law of Decreasing Credibility

Determining what can be learned using the data alone provides a logical starting point for empirical analysis, but ordinarily will not be the ending point. Having determined what can be learned in the absence of assumptions, we should then ask what more can be learned if assumptions of different strengths and degrees of plausibility are imposed.

As an econometrician, I cannot recommend that empirical researchers make one particular assumption or another—what assumptions are credible must depend on the context. My objective has been to provide a menu of possibilities. I have particularly wanted to clarify the dilemma that researchers face as they decide what assumptions

to maintain. I have called this dilemma

The Law of Decreasing Credibility: The credibility of inference decreases with the strength of the assumptions maintained.

Coping with Ambiguity

The scientific community tends to reward researchers who produce strong findings and the public tends to reward those who make unequivocal policy recommendations. These incentives tempt researchers to maintain assumptions far stronger than they can persuasively defend, in order to draw strong conclusions. We need to develop a greater tolerance for ambiguity. We must face up to the fact that we cannot answer all of the questions that we ask.

My research on identification has yielded a number of formal negative results. I have reported simple "impossibility theorems" showing that broad classes of hypotheses are not empirically testable unless sufficiently strong assumptions are maintained. I have found that researchers are often reluctant to acknowledge that, given the available data, they can logically draw some conclusion of interest only if they

maintain strong assumptions which may have limited credibility. Be that as it may, I feel strongly that both positive and negative findings contribute to the advancement of the social sciences.

Empirical Inference in Everyday Life

Social science seeks to understand the behavior of individuals and their social interactions. In their day-to-day lives, ordinary people face problems of empirical inference – problems of both identification and statistical inference – similar to those that confront social scientists. People facing inferential problems are subject to the same rules of logic as are social scientists: the conclusions that people can logically draw are determined by the assumptions and the data that they bring to bear.

Social scientists need to keep this constantly in mind as we seek to model and interpret human behavior. We do not know much about how people deal with the inferential problems that they face. Economists have been particularly negligent. Economists usually suppose that people's empirical inferences are expressed in their expectations for the future. Expectations are a subjective concept, but economists have long exercised a self-imposed prohibition on the use of subjective data in

empirical analysis. Rather than seek to measure expectations, economists have generally made assumptions about expectations.

The rational expectations assumptions commonly made by economists may be elegant and analytically appealing. However they have little empirical support. In many applications, accepting a rational expectations assumption means accepting the idea that ordinary people somehow are able to solve identification problems that have long challenged social science research. As I see it, ordinary people – like social scientists – have to cope with ambiguity.

I think that we will learn how persons cope with their inferential problems only if we probe their thought processes. One useful research activity is to measure the expectations that people hold for future events that are relevant to them. I have performed much research of this type in recent years and others have as well.

3. Application to the Selection Problem

Social scientists constantly ask "treatment effect" questions of the form: What is the effect of ____ on ____? For example, What is the effect of welfare programs on labor supply? What is the effect of schooling on wages? What is the effect of sentencing of offenders on recidivism?

Empirical analysis of treatment effects poses a fundamental identification problem, commonly called the *selection problem*. The researcher wants to compare the outcomes that persons would experience if they were to receive alternative treatments. However treatments are mutually exclusive. At most, the researcher can observe the outcome that each person experiences under the treatment that this person actually receives. The researcher cannot observe the outcomes that persons would have experienced under other treatments. These other outcomes are counterfactual. Hence data on treatments and outcomes cannot by themselves reveal treatment effects.

The Returns to Schooling

Ordinary persons want to learn treatment effects in everyday life, and so face the selection problem. Consider, for example, youth deciding whether to continue their schooling or to enter the labor market. To make good decisions, youth want to learn their returns to schooling. Youth may be able to observe the outcomes experienced by family, friends, and others who have made their own past schooling decisions. However they logically cannot observe what outcomes these people would have experienced had they made other decisions. Thus youth making schooling decisions in ordinary life are "adolescent econometricians," who face identification problems similar to those that have made it so hard for labor economists to agree on the returns to schooling.

Point-Identification of Treatment Response

Point-identification of treatment response requires assumptions about the process determining treatment selection and outcomes. The most longstanding practice, and still the most prevalent one, is to assume that among persons with specified observable covariates, treatment selection is statistically independent of outcomes.

This assumption is variously called *random*, *exogenous*, or *ignorable* treatment selection. The specified covariates are often, misleadingly, said to "control for" treatment assignment.

The assumption of random treatment selection is appropriate in the analysis of data from classical randomized experiments. Indeed this is the reason why randomized experiments are valued so highly. The assumption of random treatment selection is usually suspect in non-experimental settings, where observed treatments may be self-selected or otherwise chosen purposefully.

Over the years, a variety of alternative assumptions have been proposed and applied to non-experimental data. Indeed, the development by econometricians of selection models and instrumental-variable approaches in the 1970s was initially greeted with enthusiasm as "solving" the problem of identifying treatment effects from non-experimental data. It soon became apparent, however, that these approaches replace the suspect assumption of random treatment selection with alternative assumptions that are no less suspect.

Comparing Treatments Using the Empirical Evidence Alone

My research has moved away from the conventional focus on assumptions that yield point-identification of treatment response. I began by asking what can be learned about treatment response from the empirical evidence alone, given no assumptions about the process generating treatments and outcomes. I found that this question has a simple answer. Observation of realized treatments and outcomes does imply restrictions on the distributions of outcomes under alternative treatments. However the data are necessarily consistent with the hypothesis that there is a common distribution of outcomes under every treatment. Hence empirical evidence alone cannot determine whether one treatment is better than another.

Treatment Choice Under Ambiguity

My most recent research explores the implications of the selection problem and other identification problems for treatment choice. I suppose that a social planner must choose a treatment rule assigning a treatment to each member of a heterogeneous population. The planner could, for example, be a physician choosing medical treatments for each member of a population of patients, a school official making

course placement decisions for each member of a population of students, or a judge deciding sentences for each member of a population of convicted offenders.

I suppose that the planner observes certain covariates for each person. These covariates determine the set of treatment rules that are feasible to implement: the set of feasible rules is the set of all functions mapping the observed covariates into treatments. Each member of the population has a response function mapping treatments into real-valued outcomes. I suppose that the planner wants to choose a treatment rule that maximizes a utilitarian social welfare function.

Suppose for simplicity that all treatments have the same costs. Then it is easy to show that an optimal treatment rule assigns to each member of the population a treatment that maximizes mean outcome conditional on the person's observed covariates. The planner faces a problem of treatment choice under *uncertainty* if he knows the conditional mean responses and, consequently, can implement an optimal rule. The planner faces a problem of treatment choice under *ambiguity* if he does not know enough about mean response to be able to implement an optimal rule.

Identification problems make ambiguity a fundamental problem of treatment choice

in practice. Although empirical evidence on realized treatments and outcomes does imply informative bounds on mean responses under alternative treatments, these bounds may overlap. Then observations of realized treatments and outcomes do not suffice to rank the feasible treatment rules. This does not imply that a planner should be paralyzed, unwilling and unable to choose a rule. There are various reasonable ways to proceed, but not one best way. Thus, planners must somehow cope with ambiguity. They cannot simply assume it away.

So must ordinary persons cope with ambiguity as they face their own decision problems. How do we cope? This is the question that most intrigues and perplexes me as I look ahead to my future research.